基于自学习近邻图策略的短文本匹配方法 *

付 聪，李六武，杨振国，刘文印

(广东工业大学 计算机学院，广州 510006)

摘 要：针对自然语言处理中的文本匹配问题，提出一种基于自学习文本近邻图框架的深度学习模型，以处理短文本匹配问题。文本近邻图可使用词嵌入将文本转换为向量形式，再通过构建文本相似度关系矩阵获得，可表达文本样本的近邻关系。现有方法通常构造静态的近邻图，这些方法一方面依赖先验知识，另一方面难以获得句子对的最优表示。因此，提出了利用孪生卷积神经网络学习更优的动态更新的近邻图。该模型在 Quora 数据集上的准确率和 F1 值分别是 84.15%和 79.88%，在 MSRP 数据集上的准确率和 F1 值分别是 74.55%和 81.63%。实验表明，提出的模型能有效地提高文本识别和匹配的准确率。

关键词：文本匹配；自学习近邻图；词嵌入；孪生卷积神经网络

中图分类号：TP391.1 doi: 10.19734/j.issn.1001-3695.2018.12.0877

Self-adaptive affinity graph learning for short text matching

Fu Cong, Li Liuwu, Yang Zhenguo, Liu Wenyin

(*School of Computer Science*, *Guangdong University of Technology*, *Guangzhou* 510006, *China*)

Abstract: For text matching problems in natural language processing, this paper proposed a deep learning model based on self-adaptive affinity graph learning framework for short text matching. The affinity graph can be converted into a vector form using word embedding, and then obtained by constructing a text similarity relationship matrix, which can express the neighbor relationship of the text sample. Current methods usually construct static affinity graphs, which rely on prior knowledge and hard to obtain the optimal representation of sentence pairs. Therefore, this paper proposed to use the Siamese CNN to learn the affinity graph of better dynamic updates. The accuracy and F1 values of the model on the Quora dataset are 84.15% and 79.88%, respectively, and the accuracy and F1 values on the MSRP dataset are 74.55% and 81.63%, respectively. Experiments show that the proposed model can improve the accuracy of text recognition and matching effectively.

Key words: text matching; self-adaptive affinity graph learning; word embedding; siamese CNN

0 引言

近年来，在互联网的高速发展下，网络上的信息可以被人们越来越方便地获取到，由于网络信息的爆炸性增长，使得文本数据越来越多，因此从这样巨大的知识库里搜索到自己需要的信息便成为一个富有挑战性的任务[1]。在这种需求下，社区问答（community question answering, CQA）系统应运而生，逐渐成为一种非常受欢迎的互联网应用，如雅虎问答、新浪爱问和百度知道等。用户可以在问答社区提交问题，也可以回答其他用户提出的问题。现代信息检索系统虽然已经能够基本满足用户对于信息查找的需求，但在检索过程中并未直接提供有用的相关信息，用户需要阅读大量相关文档才能找到自己需要的信息。因此目前的搜索引擎并不能完全满足用户对搜索质量的高要求。如何从这样庞大的信息网络中搜索到对自己有用的信息？如何让信息搜索的效率和质量更高？针对上述问题研究发现，对用户提出的问题进行相似性检测，进而把最相关的优质答案推送给用户是解决这类搜索问题的有效方法。随着深度学习方法在自然语言处理（natural language processing,NLP）应用方面越来越受欢迎，深度学习技术在解决自然语言处理相关问题中取得了越来越多的突出表现[2]。对于问答系统文本相似性的检测，本文提出了基于深度学习的 AutoLMP（auto-learning match pyramid）模型，采用自学习方式生成的文本内容信息图和 Pang 的方法得到的文本相似度图共同组成一幅表达丰富的二通道近邻图。这两个通道上一一对应的像素值都是来自相同的两个单词向量，一个像素值代表由 SCNN 学习到的文本内容信息，另一个像素值代表通过先验知识计算出来的文本相似度信息；然后经过层级 CNN 提取出句子中丰富的语义信息和词与词之间的关系，层级 CNN 能够从词到句子水平提取出更复杂更丰富的匹配信息；最后在相似问题匹配和释义识别两个任务上进行实验，结果表明采用的方法表现出了良好的效果。

1 相关理论

1.1 自然语言处理与卷积神经网络

随着词向量[3]、分布式特征表示和神经网络语言模型[4]等的兴起和发展，深度学习方法在自然语言处理领域发挥了越来越重要的作用，这里主要对 CNN 在 NLP 领域的应用做一个详细的介绍。由于 CNN 在空间上共享参数，从而减少

收稿日期：2018-12-08；**修回日期**：2019-02-12 **基金项目**：国家自然科学基金资助项目（61703109，91748107）；中国博士后科学基金资助项目（2018M643024）；广东省引进创新科研团队计划资助项目（2014ZT05G157）

作者简介：付聪（1991-），女，山东济宁人，硕士研究生，主要研究方向为自然语言处理、文本挖掘；李六武（1992-），男，广西梧州，硕士研究生，主要研究方向为计算机视觉、自然语言处理；杨振国（1988-），男，山东潍坊，博士（后），主要研究方向为自然语言处理、文本挖掘、多媒体；刘文印（1966-），男，吉林榆树人，教授，硕/博导，主要研究方向为文本挖掘、区块链、网络身份安全等（liuwy@gdut.edu.cn）.

了神经网络中参数的个数。CNN 通过多层训练的网络空间结构[5]，不仅在很大程度上减少了参数量，而且还提高了训练效率，避免了全连接网络因为参数过多不好训练以及梯度弥散的问题。此外，在文本分类中 CNN 模型也取得了不错的效果,该模型最初是为计算机视觉而发明的，后来被证明对 NLP 有效，并且在语义分析、查询检索[6]、句子建模等方面都取得了优异的成果。众所周知，CNN 模型一开始是普遍应用在图像领域的，经过预处理，每一个图像的高和宽具有相同的像素值，之后对该图像做卷积操作。但文本和图像的处理方法是不一样的，由于在文本语料中句子的长度是不固定的，就需要把它处理成和图像的二维矩阵类似的结构才能够进行实验，一方面每个句子应该扩充到最大句子长度，另一方面使用词嵌入直接在神经网络中从头开始训练词向量，词向量的训练可以使用 FastText[7]或者 word2vec[3]等方法，把训练好的词向量作为神经网络中嵌入层的权重，之后进行微调。Kim[8]应用了 CNN 的经典构造，将长度不同的过滤器在文本矩阵上做卷积操作，文本矩阵中词向量的长度与过滤器的宽度相同；接着将提取出的向量运用最大池化进行实验；最后把所有过滤器对应的相应数字拼接起来，就能得到对应的句子向量。根据 CNN 在 NLP 应用的这些特点和优势，本文也使用 CNN 来搭建模型。

1.2 词嵌入

词嵌入在 NLP 中应用广泛，该技术能够把词语和文本转换成计算机可以处理的向量形式，这是文本处理的第一步。依据目前的发展，词的向量化表示分为以下三种：

a) 独热表示是过去比较常用的表示方法，通过该方法每个词被表示为一个维数很高的向量，向量的维度代表了词表的大小，每个词向量的数值里只有一个维度的值为 1，其余全为 0，当下的词就由这个 1 来表示，因此这种方法表示的词向量就很稀疏。除此之外，这种表示方法的不足之处还有：(a)词与词之间的关系都是独立的，对于语义关系相同或相似的信息无法表达；(b)句子中词的种类数决定了向量维度的大小，在很多情况下这种表示会导致词典变得很大；

b) 词的分布式表示在一定程度上克服了这个缺点，这种表示可以把词映射到相对低维、密集的向量空间里，将词符号化之后，利用向量公式来计算词之间的相似性，对于上下文比较相似的词，其对应的语义也是类似的。

c) word2vec 是 2013 年由 Mikolov 等人[3]提出的，这种表示方法能够有效地降低词向量的维度，分为两种模型：一种是 Skip-gram 模型，它是通过输入某个词语来预测该词语的上下文;另一种是连续词袋 (continuous bag-of-words，CBOW) 模型 ,该模型是从上下文对目标词的预测中学习到词向量的表达（即输入上下文来预测当前词），本文使用的就是 CBOW 模型。

1.3 文本相似度计算

问题相似性检测的核心是文本相似度计算。在自然语言处理中，文本相似性分析是一项重要且具有挑战性的任务。近年来，深度学习模型在语音识别[9]和计算机视觉等许多领域都取得了不错的效果。在 NLP 领域中，基于深度学习的多种模型设计和方法随后也蓬勃发展了起来[10]。Huang 等人[11]提出一种经典的单语义模型 DSSM（deep structured semantic models）,该模型是一个具有深层结构的潜在语义模型，可以将查询和文档投影到一个共同的低维空间中；文中还使用了一种称为单词散列的技术，该技术不仅可以有效地扩展语义模型，而且还能让模型适用于大规模 Web 搜索应用程序。但是该模型也有不足之处，比如在做相似度匹配任务时，使用的是无参的余弦相似度匹配公式；而且还忽略了单词之间时序关系。Mueller 等人[12]提出了一种基于孪生递归神经网络的学习文本相似度的模型体系结构，用于学习变长字符序列的相似性度量。该模型将一堆字符级双向长短期记忆网络与孪生架构相结合，通过使用有关字符串对之间相似性的信息，学习将可变长度字符串投影到固定维度嵌入空间中；但是在职称标准化的任务中，该模型应用的是基于手动注释的分类法，比较费时费力。Pang 等人[13]提出了 Match Pyramid 模型，该模型将文本匹配作为图像识别，把图像识别中的卷积神经网络思想迁移到了文本匹配中，通过匹配矩阵捕获不同层次的匹配模式，从单词、短语到整句话，论文主要思想就是将文本匹配建模为图像识别，将匹配矩阵作为图像。由于这种方法是构造静态的近邻图，一方面依赖先验知识，另一方面难以获得句子对的最优表示。鉴于这个缺陷，本文采用 SCNN 深度学习模型，提出一种借助 SCNN 生成可学习文本内容矩阵的方法，以此捕获文本中的关键性信息，从而更好地识别和检测问答系统的相似问题。

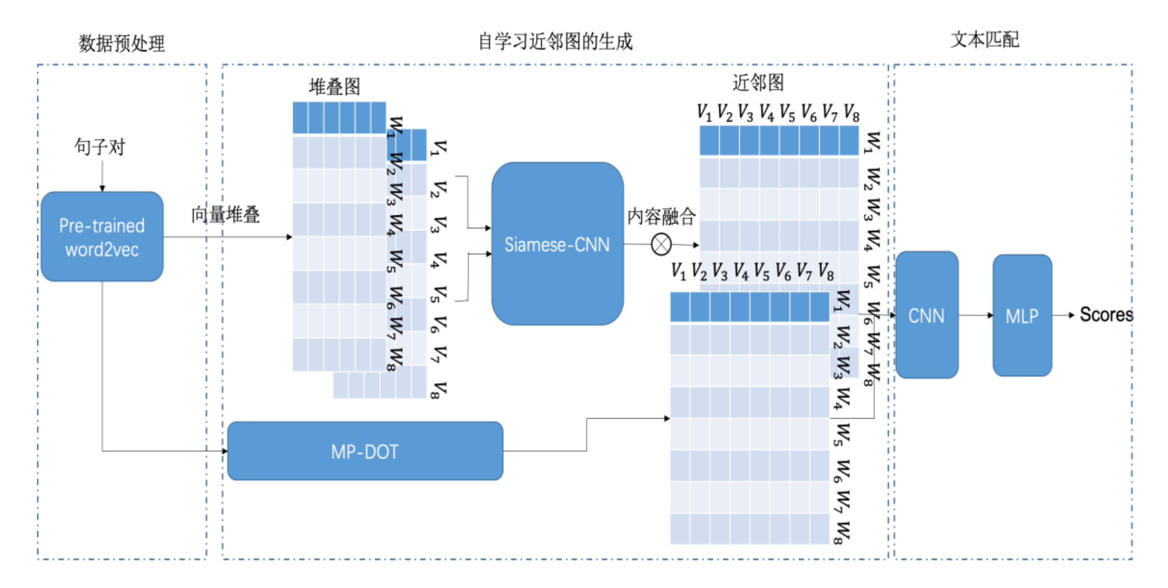


图 1 模型结构：auto-learning MatchPyramid

Fig. 1 Model structure : auto-learning MatchPyramid

2 AutoLMP 模型

本文提出一种新的深度文本匹配框架，称之为 AutoLMP（auto-learning match pyramid）。本文模型主要的灵感来自图像识别，通过将文本映射为一张图片来构建基于卷积神经网络的文本匹配。总体框架如图 1 所示。

本文的 AutoLMP 框架主要由三部分组成：a)数据预处理，这部分将句子对向量化，为近邻图的生成做好数据准备；b)本文提出的近邻图的生成过程，其中包括自学习近邻图的生成过程和先验知识近邻图的生成过程；c)文本匹配，包括层级 CNN 和多层感知器。

2.1 构建近邻图

一个句子对和一幅图像有许多相似之处，比如：a)句子的基本构成元素是单词，图像的基本构成元素是像素；b）句子对中单词与单词有着或多或少的关系，在图像中像素与像素之间也有着千丝万缕的关系；c）句子表达着某种明确的信息，图像也展现着一个明确的场景。因此，图像识别的方法通过模式的转换用到文本匹配任务上。受 Pang 提出的 Match Pyramid[13]方法的启发，文本匹配任务可以通过构建近邻图来使用深度卷积神经网络来完成。其中，近邻图是匹配句子对的关系矩阵。例如相似度矩阵，它包含了句子对之间的丰富信息。因此，构建近邻图是使用深度卷积神经网络来解决文本匹配的关键步骤。构建文本关系矩阵有许多方法，如 Match Pyramid 中计算句子对中各个词之间的 cosine 值。但是通过这些方法所得到的近邻图都是经过先验知识计算出来的，近邻图所包含的信息并不代表是句子对的最优表示，而神经网络却可以通过学习来获得更优的近邻图，因此本文使用神经网络的结构来构建模型。下面将详细介绍提出的 AutoLMP 模型。

2.2 AutoLMP 模型细节

a）数据预处理。在进行近邻图的生成之前，本文首先将句子对分别词向量化，采用两百万条 Twitter 消息预训练的 word2vec 模型 Glove 生成的 50 维向量；然后将每个句子的单词向量纵向堆叠，纵向维度取句子的最长长度为 200，长度不足的用 0 补充，这样就可以得到一幅 200*50 的堆叠图。

b）文本匹配。如图 2 所示，字级匹配指的是两个文本中单词之间的匹配，不仅包括相同的单词匹配，如 Comedy-Comedy、Nights-Nights、with-with、Kapil-Kapil、live-live,还包括类似的词匹配，如 watch-see；短语级别匹配指的是短语之间的匹配，即 N-gram 匹配，指的是 n 个连续单词发生的匹配，如(What is the way)-(How can)、live on the sets-live show；句子级别匹配指的是句子之间的匹配，由多个较低级别的匹配单元组成，如上面的这一对句子可从单词和短语层面进行匹配。当考虑包含多个句子的段落之间的匹配时，整个段落将被视为一个长句子。



图 2 文本不同层级匹配

Fig. 2 Different level of text mathing

c）SCNN 结构。正如前面所论述，近邻图的构建是将文本匹配问题转换为图像识别问题的关键。一幅近邻图是二维的像素矩阵，而每个像素都与句子对的单词一一对应，相当于将有序的单词信息映射到结构性的二维矩阵上。为了解决这个问题，本文用 M 表示近邻图。M 可以从本文构建的可学习的 SCNN 生成，可以用式（1）表示，其中：g 和 g′ 分别表示由句子对分别生成的堆叠图（stacking gragh）；$\overrightarrow{w_i}$和$\overrightarrow{v_j}$分别表示 g 和 g′ 通过 SCNN 学习到的句子对向量；而近邻图 M_{ij} 则由 $\overrightarrow{w_i}$和$\overrightarrow{v_j}$ 点乘得到。下面本文将呈现更多的细节。

$$\overrightarrow{w_i}, \overrightarrow{v_j} = \mathrm{SCNN}(g, g') \tag{1}$$

$$M_{ij} = \overrightarrow{w_i}^T \overrightarrow{v_j} \tag{2}$$

本文构建的 SCNN 如图 3 所示。它是由三层全卷积神经网络组成，每层卷积层的输出都进行归一化，并且使用 Relu 激活函数激活。SCNN 共享参数，一次处理一个句子对，使得句子对可以使用相同的函数映射规则，这样做使得得到的句子都具有一致性。

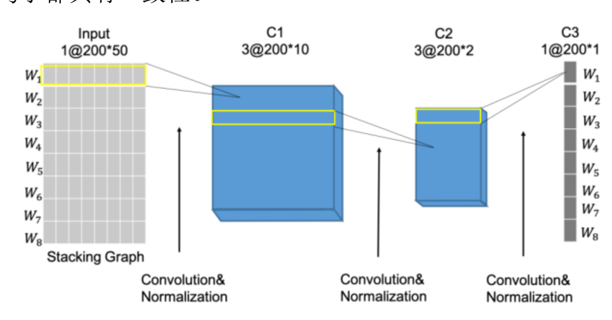


图 3 SCNN 整体结构

Fig. 3 Overall structure of SCNN

d）SCNN 的内部参数设置。如图 4 所示，三层卷积层的卷积核大小分别是 1*5、1*5 和 1*2，步长分别是 5、5 和 1，卷积核个数分别是 16、16 和 1。一幅 200*50 大小的堆叠图经过 SCNN 之后得到 200*1 的向量。本文的卷积层这样设置的意义有：(a)由于堆叠图每一行代表一个单词的向量，1*N 的卷积核只学习单词向量内的信息，并不会引入相邻单词向量的噪声信息，如图 4 所示单词向量 W_i 经过 SCNN 之后仍然与单词向量 W_i 对应；(b)步长的设置可以保证每层学习的信息不一样，减少冗余信息。另外，合适的卷积核保证足够的模型学习能力。

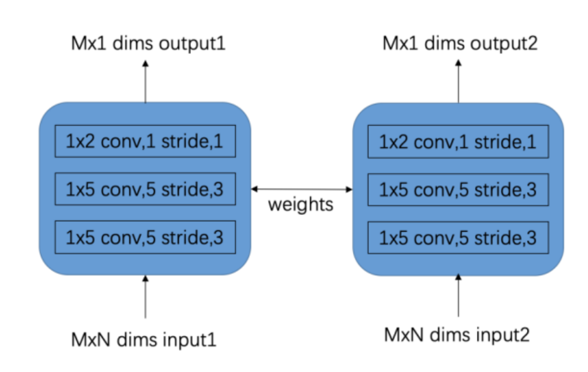


图 4 SCNN 的内部结构

Fig. 4 Internal structure of SCNN

e）自学习近邻图的生成。如图 5 所示，代表每个句子对的堆叠图经过 SCNN 学习得到其对应的向量，接着这其中一个向量转置之后与另一个向量做点乘即得到该句子对的近邻图 M_{ij}。如图 5 所示，单词向量 W_i 与 V_j 经过 SCNN 后被映射到近邻图 M_{ii} 的位置。近邻图 M_{ii} 上的每个像素都代表着句子对中每两个单词的内容信息，也就是说本文构造的 SCNN 学习的近邻图是代表着句子对的完整内容信息。

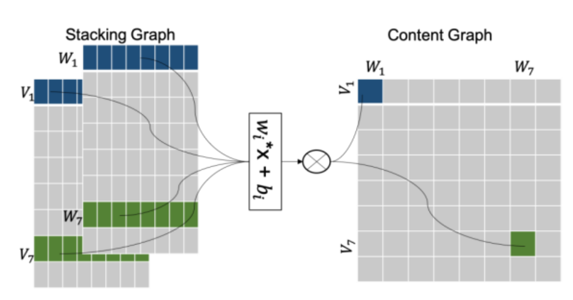


图 5 自学习内容信息矩阵生成

Fig. 5 Self-learning content information matrix generation

f）先验知识近邻图的生成。本文积极采用 Pang 的方法，由于在其生成近邻图的方法中最好的是向量点积方法，所以本文也使用该方法得到另一张近邻图。本文用自学习方式生成的文本内容信息图和采用 Pang 的方法得到的文本相似度图共同组成一幅表达丰富的二通道近邻图。更重要的是，这两个通道上一一对应的像素值都是来自相同的两个单词向量，而且一个像素值代表学习到文本的内容信息，另一个像素值代表通过先验知识计算来的文本相似度信息。这是本文提出的 AutoLMP 模型的本质。

2.3 AutoLMP 模型的关键方法—层级 CNN

下面将详细剖析本文使用的文本匹配层级 CNN 方法。该方法可以提取包括单词、词组、句子等不同水平的匹配模式。对于层级 CNN 的第一层，第 k 个卷积核 $w^{(1,k)}$ 在双通道近邻图上依次滑动做卷积计算来产生一张特征图 $m_{i,j}^{(1,k)}$：

$$m_{i,j}^{(1,k)} = f\left(\sum_{p=0}^{g_k-1}\sum_{q=0}^{g_k-1} w_{p,q}^{(1,k)} * m_{i+p,j+q}^{(0)} + b^{(1,k)}\right) \tag{3}$$

其中：g_k 表示第 k 个卷积核的大小。在本文中使用 n*n 形式的卷积核和 ReLU 激活函数。ReLU 的公式为

$$R(x)=\begin{cases}0 & x<0\\ x & x\ge 0\end{cases} \tag{4}$$

另外，由于句子长度大小不一，本文采用动态池策略来解决这个问题。在使用动态池策略得到的特征图可以表示为

$$m_{i,j}^{(2,k)} = \max_{0\le p<d_k}\max_{0\le q<d_k'} m_{i\cdot d_k+p,\, j\cdot d_k'+q}^{(1,k)} \tag{5}$$

式（4 中）：d_k 和 d_k' 分别表示动态池对应的池化核的长和宽（$d_k = x/x'$，$d_k' = y/y'$），池化核的长和宽由句子对长度 x 和 y 决定。经过池化得到的特征图大小分别为 x'*y'。

在经过第一层卷积和动态池化后，将得到更高水平的特征图 $m^{(l)}$，当 l>2 时，层级 CNN 继续加深层数，之后的卷积核最大池化可以泛化表示为

$$m_{i,j}^{(l+1,k')} = f\left(\sum_{k=0}^{c_l-1}\sum_{p=0}^{g_k-1}\sum_{q=0}^{g_k-1} w_{p,q}^{(l+1,k')} * m_{i+p,j+q}^{(l,k)} + b^{(l+1,k)}\right) \tag{6}$$
$$l = 2,4,6,\ldots$$

$$m_{i,j}^{(2,k)} = \max_{0\le p<d_k}\max_{0\le q<d_k'} m_{i\cdot d_k+p,\, j\cdot d_k'+q}^{(l,k)} \tag{7}$$
$$l = 3,5,7,\ldots$$

其中：c_l 表示第 l 层的特征图个数。

2.4 对层级 CNN 有效性的分析

CNN 在图像识别中能够有效地提取图片的基本视觉元素，如边和角。在本文提出的 AutoLMP 模型中，层级 CNN 也能够提取丰富的语义信息和词与词之间的关系。如图 6 所示，本文将举例展示层级 CNN 是如何进行特征提取的。

a)在双通道的近邻图上，格子 M_{ij} 表示句子 1 第 i 个单词与句子 2 第 j 个单词的映射结果。如图 6 所示，第一层卷积层的两个 3*3 的卷积核将句子对中相邻的三个单词的两两关系依次被映射到更高层次的特征图上。这种模式就像卷积核提取图像的边缘特征一样。

b)从第一层得到的多张特征图会被下一层卷积层继续提取更高维的特征。比如这一层依然采用 3*3 的卷积核去提取特征图上的特征，由于这些特征图的每一格表示句子对中相邻的三个单词的两两关系的映射，现在每次处理的是句子对中相邻的九个单词的关系信息，这样更多的句子信息和单词信息被提取到，就像图像识别中更深层的网络提取到更抽象的图像特征。

从上面的分析中可以看到，层级 CNN 能够从词到句子水平上提取更复杂更丰富的匹配信息。

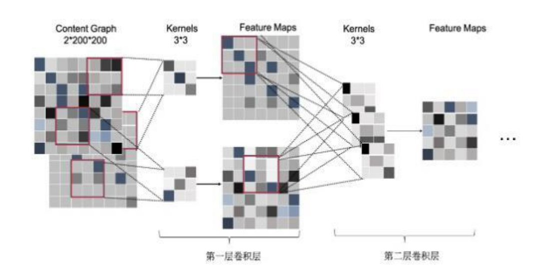


图 6 层级 CNN 提取特征过程分析

Fig. 6 Analysis of hierarchical CNN extraction feature

2.5 预测网络和目标函数

在 AutoLMP 模型中，本文使用了多层感知器去预测匹配分数，然后从多层感知器得到一个二分类结果。现在用两层感知器来说明感知层是如何计算的：

$$s = W_2\delta(W_1 m + b_1) + b_2 \tag{8}$$

其中：s 是句子对的匹配概率；m 是层级 CNN 的输出；W_i 是第 i 层感知器的权值；δ 表示激活函数。

在 AutoLMP 模型中，本文使用 Sigmoid 函数处理模型的输出来产生句子对的类别预测概率，然后使用二分类交叉熵 loss 作为目标函数来训练。最后优化使损失降到最小，即

$$\text{loss} = \sum_{i=1}^{N}[l^{(i)}\log(s_i)+(1-l_i)\cdot\log(1-l_i)] \tag{9}$$

其中：l_i 是第 i 个句子对的标签；s_i 是第 i 个句子对的预测概率。

3 实验

在本章中对两个任务进行实验，即相似问题匹配和释义识别，以证明 AutoLMP 比基线的优越性。

3.1 数据集

1）Quora 重复问题数据集

作为一个受欢迎的知识分享平台，Quora 中类似的问题不应该多次出现。本文采用的就是 Quora 公开的问题数据集，共有 40 万+的问题对。检测一个问题对在语义上是否相似，就是把两个问题 q1 和 q2 作为输入，经过模型处理输出结果，根据输出的概率值来判定这两个问题在语义上是否存在相似关系，结果越接近 1 表示问题对越相似。在数据集中有大约 63％的非重复问题和 37％的重复问题。

2）MSRP 数据集

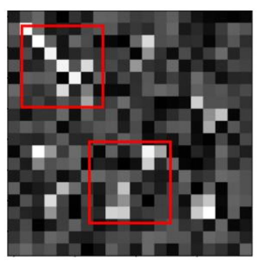
为了进一步验证本文提出的 AutoLMP 模型的泛化性和有效性，本文还使用基准 MSRP 数据集进行测试。该数据集包含 4 076 个用于训练的实例和 1 725 个用于测试的实例，意在判别两个句子是否具有相同的意思，其属于文本匹配的范畴。

3.2 方法有效性论证

如图 7 所示，两个矩阵分别是 Pang 的方法生成的匹配矩阵和本文 AutoLMP 方法生成的自学习内容信息矩阵。通过

对比两个矩阵发现，自学习内容信息矩阵与匹配矩阵一样具有一定的特征分布。正如左、右四个红色方框圈的区域，匹配矩阵和自学习内容信息矩阵都拥有较多的白块，也就是匹配矩阵具有的特征，自学习内容信息矩阵也一样拥有，

而且自学习内容信息矩阵拥有更多的特征信息。综合以上，本文提出的 AutoLMP 方法生成的自学习内容信息矩阵能够表示句子对的内容信息特征，而且与匹配矩阵形成了近邻关系，两个句子组成的近邻图能够表达丰富的句子对信息。经过多次实验发现，由于句子对的不同，自学习内容信息矩阵所具有的特征分布变化较大，但总体能体现一定的特征。



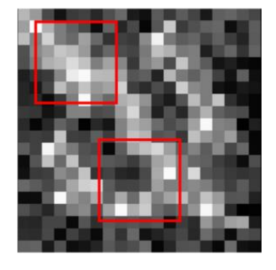


图 7　匹配矩阵和自学习内容信息矩阵

Fig. 7　Matching matrix and self-learning content information matrix

3.3　方法对比

为了验证 AutoLMP 模型的有效性，本文用多种方法与 AutoLMP 模型进行实验对比。以下是本文使用到的模型，首先对实验数据集进行预处理，包括分词、词干化、文本大小写转换、去掉停用词等；之后对归一化的数据集进行训练，基于需要完成的任务，AutoLMP 自动学习滤波器的权重值。针对文本矩阵图，为了检验生成不同文本矩阵方法的有效性和可靠性，本文设置了两组实验，每组实验各采用不同的经典方法如 TF-IDF, DSSM, ARC-I, MP-DOT 等，与本文提出的模型 AutoLMP 进行比较实验。以下是各方法介绍：

a）TF-IDF。TF-IDF 是文本挖掘中广泛使用的方法。在这种方法中，每个文本都表示为|V|维向量，每个元素代表文本中相应单词的 TF-IDF 得分，其中|V|是词汇的数量。在本文实验中，IDF 得分是在整个数据集中计算的，最终匹配分数由两个向量的内积产生。

b）DSSM / CDSSM。由于 DSSM 和 CDSSM 需要大量数据进行训练，本文直接使用已发布的模型，通过大型数据集来训练本文的测试数据。

c）ARC-I / ARC-II。本文使用 ARC-I 和 ARC-II[14]，由于没有公开的代码，所以此方法使用的是与原始论文相同的设置。

3.4　实现细节

本文通过多轮的测试，最终确定了超参数的设置；训练过程中误差传递采用随机梯度下降（SGD）的方法，学习率设置为 0.1；对于误差的更新，则采用批处理的形式，每次由 64 个样本一起更新，即 batch 设置为 64 个，drop-out 设置为 0.5，迭代轮回次数设置为 500，选择在评估数据集上性能最好的一组参数作为训练的模型参数输出。

3.5　相似问题匹配评测

实验结果如表 1 所示。可以看到，在传统方法中，像 ARC-II 方法最高已经获得 80%以上的准确率，尽管这些传统方法仅使用比较低级的文本特征信息。通过先验知识获得近邻图的 MP-DOT 方法在该测试集的实验结果有了较大的提升，其获得 83.48%的准确率和 79.37%的 F1 值。另外采用本文提出方法中的 Single AutoLMP 方法（仅采用自学习得到的单通道近邻图）也获得 82.77%的准确率和 78.96%的 F1 值，而采用本文提出的整体方法 AutoLMP 获得了最好的效果，其中，准确率达到了 84.15%，F1 值达到了 79.88%。实验结果表明，将文本匹配任务建模为图像识别是一个很好的解决方案。结果也进一步证明了提出的 AutoLMP 方法的优越性，本文方法既利用了通过先验知识获得的近邻图，又利用了通过学习得到的近邻图。双通道近邻图包含了更丰富的文本特征信息，采用层级 CNN 能够有效地提取文本特征信息。

表 1　Quora 上的结果

Table 1　Results on Quora

Model	ACC(%)	F1(%)
TF-IDF	78.59	69.22
DSSM	68.53	28.97
CDSSM	65.78	18.68
ARC-I	79.96	74.62
ARC-II	80.14	75.93
MP-DOT	83.48	79.37
SingleAutoLMP	82.77	78.96
AutoLMP	**84.15**	**79.88**

3.6　释义识别

实验结果列于表 2。可以看到，传统的简单模型如 TF-IDF 已经达到了约 69.32%的高精度，尽管它只使用单字母匹配信号。本文方法比 TF-IDF 表现得更好，这表明层级卷积捕获的复杂匹配模式对文本匹配任务很重要。通过与近期深度模型的比较，可以看出本文模型（AutoLMP）表现优于所有模型。在层级 CNN 方法基础上，无论利用单通道的 MP-DOT，还是利用单通道的 Single AutoLMP，都获得了比传统方法更优的实验结果。虽然利用本文提出的双通道的 AutoLMP 方法相对于前作 MP-DOT 方法仅有一点提高，但是这也证明该方法是可行的方法，仍然有很大改进的空间。在未来的工作中，本文将更深入研究自学习近邻图生成来进一步改进本文的模型。

表 2　MSRP 上的结果

Table 2　Results on MSRP

Model	ACC(%)	F1(%)
TF-IDF	69.32	75.89
DSSM	68.87	79.03
CDSSM	66.89	78.94
ARC-I	66.35	78.62
ARC-II	66.41	78.55
MP-DOT	73.90	80.92
SingleAutoLMP	73.06	79.27
AutoLMP	**74.55**	**81.63**

4　结束语

本文提出了基于自学习近邻图的有效层级 CNN 方法，将文本信息编码为词向量以后，通过可学习的 SCNN 得到近邻图，再结合先验知识得到 Matching Matrix，从而得到语义信息更全面的双通道文本匹配矩阵。通过多个对比实验，可以更好地对相似文本进行匹配，证明了本文方法的可行性和有效性。下一步工作将进一步探讨文本更多层级和粒度的语义分析任务，寻找更适合文本语义相似性分析的深度学习算法。

参考文献：

[1] 黎新. 面向问答系统的段落检索技术研究 [D]. 合肥：中国科学技

术大学, 2010. (Li Xin. Research on paragraph retrieval technology for question answering system [D]. Hefei: University of Science and Technology of China, 2010.)

[2] 林奕欧, 雷航, 李晓瑜, 等. 自然语言处理中的深度学习方法及应用 [J]. 电子科技大学学报, 2017, 46 (6): 913-919. (Lin Yiou, Lei Hang, Li Xiaoyu, *et al.* Deep learning method and application in natural language processing [J]. Journal of University of Electronic Science and Technology of China, 2017, 46 (6): 913-919.)

[3] Mikolov T, Kai Chen, Corrado G, *et al.* Efficient estimation of word representations in vector space [J/OL]. Computer Science, 2013. arXiv:1301.3781v3

[4] Wei Xu, Rudnicky A. Can artificial neural networks learn language models? [C]// Proc of the 6th International Conference on Spoken Language Processing. Beijing: ICSLP Press, 2000.

[5] Lecun Y L, Bottou L, Bengio Y, *et al.* Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.

[6] Shen Yelong, He Xiaodong, Gao Jianfeng, *et al.* Learning semantic representations using convolutional neural networks for Web search [C]// Proc of International Conference on World Wide Web. 2014: 373-374.

[7] Joulin A, Grave E, Bojanowski P, *et al.* Bag of tricks for efficient text classification [J/OL]. Preprint arXiv 2016: 1607. 01759.

[8] Kim Y. Convolutional neural networks for sentence classification [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2014: 1408.

[9] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2013: 6645-6649.

[10] Young T, Hazarika D, Poria S, *et al.* Recent trends in deep learning based natural language processing [J]. IEEE Computational Intelligence Magazine, 2017, 13 (3): 55-75.

[11] Huang Posen, He Xiaodong, Gao Jianfeng, *et al.* Learning deep structured semantic models for Web search using clickthrough data [C]// Proc of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2013: 2333-2338.

[12] Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity [C]// Proc of AAAI Conference on Artificial Intelligence. 2016: 2786-2792.

[13] Liang Pang, Yanyan Lan, Jiafeng Guo, *et al.* Text matching as image recognition [J/OL]. CoRR, 2016: abs//1602. 06359.

[14] Hu Baotian, Lu Zhengdong, Li Hang, *et al.* Convolutional neural network architectures for matching natural language sentences [C]//Proc of the 27th International Conference on Neural Information Processing Systems. Cambridge, MA :MIT Press, 2014: 2042-2050.